DOCUMENT RESUME

ED 301 586 TM 012 460

AUTHOR Taleporos, Betsy; And Others
TITLE The Process of Customization of the "Metropolitan Achievement Test" (MAT-6) in Mathematics for New York City Public School Students.
PUB DATE 88
NOTE 12p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Difficulty Level; Elementary Education; Elementary School Mathematics; *Elementary School Students; *Mathematics Tests; *Norm Referenced Tests; Psychometrics; Public Schools; School Districts; *Test Construction; Testing Programs; Test Items; Test Reliability; Urban Education
IDENTIFIERS Curriculum Based Assessment; *Metropolitan Achievement Tests; New York (New York Metropolitan Area); *Test Customization

ABSTRACT

In the spring of 1986, New York City began using the Metropolitan Achievement Test-6 (MAT-6) series to assess achievement in mathematics as part of a continuing end-of-year testing program. During the first two years of the program, appropriate levels of the shelf version of MAT-6 (Forms L and M) were administered to second through eighth graders. Since then, a customization process, described in this paper, has been undertaken due to concern expressed by teachers and district-level staff that the tests were not uniquely geared to assess the performance objectives specified in the New York City curriculum guides. For grades 2 through 8, approximately 200 new items were developed that clearly measured city curriculum objectives; approximately half of these items were taken from other tests that the MAT-6 publisher had available. The items were arranged on tryout sheets and piloted with student samples. The attempt to select tests that were perfect psychometric replicas of the shelf tests while also being perfect matches to the curriculum was not completely feasible. It also proved troublesome to determine difficulty levels for various items. Item selection is now completed, but equating techniques must be applied following the next spring's test administration. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED301586

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

JANE CANNER

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

The Process of Customization of the Metropolitan Achievement Test (MAT-6) in Mathematics for New York City Public School Students

Betsy Taleporos
Jane Canner
Irene Strum
Doreen Faulkner

Office of Educational Assessment
N.Y.C. Board of Education

TM 012 460

BEST COPY AVAILABLE

In New York City, all students traditionally take part in mandated, survey, end-of-year achievement tests in reading and mathematics. Although each of the 32 local districts that make up the city are free to do their own additional testing for their own needs, we strive to make the required testing program one that will meet both central and local needs, in the hope of eliminating additional testing for students. Since the citywide test is used for multiple purposes (e.g., overall assessment of our students' achievement, school-level instructional planning, curriculum planning, and program evaluation), much care goes into our test selection and/or test development process.

In the spring of 1986, after a lengthy test selection process, New York City began using a new test, the Metropolitan Achievement Test-6 (MAT-6) series, to assess achievement in mathematics as part of our continuing end-of-year testing program. In the first two years of this program, appropriate levels of the shelf version of MAT-6 (Forms L and M) were administered to students in Grades 2 through 8.

When the shelf tests were administered, teachers and local district-level staff expressed much concern that these tests were not uniquely geared to assess the performance objectives specified in the New York City curriculum guides. In particular, since some new curriculum areas that had been emphasized repeatedly in staff development were only minimally if at all included on the test (e.g., locating missing information, probability), people became quickly disenchanted with the

test, and lost sight of the fact that, overall, it represented a relatively good match to the curriculum (particularly when compared to other standardized test series examined). It nonetheless was perceived as having limited usefulness as a diagnostic and prescriptive tool for classroom teachers and curriculum program planners. In fact, the concerns voiced were so strong that the credibility of the program and the hoped-for meaningful use of test results by local districts were at stake. Many of our constituencies were calling for either a new test, a locally-developed test, or a customized version of the MAT-6; whichever option was selected, they wanted a test that would clearly be seen as assessing our N.Y.C. curriculum. Our concern was that the option chosen would allow us to continue to obtain valid norm-referenced information, which we needed. Since our initial contract with the publisher called for the possibility of a customized test, we proceeded with plans to customize the test rather than to develop our own test or to look (again) for another. The customization began, then, in response to curriculum needs, but was undertaken with a view to preserving norm-referenced information that was essential to meet a variety of legal and public reporting obligations.

The present paper represents an attempt to describe the customization process and to delineate some of the problems encountered. It is hoped that this presentation will be useful to school systems that are considering customization of shelf

tests, and add some practical experience to the growing body of literature arguing both for and against the concept of customization in large school systems.

An essential ingredient contributing to whatever success we may have had in this endeavor was that we were constantly vigilant to the cooperative nature of the project, trying at each point in time to balance psychometric considerations with those expressed by the staff of the Math Unit of our Division of Curriculum and Instruction. At the onset, we agreed that the project would strive to be a "perfect union" of psychometric and curriculum interests. We sought to develop a test that assessed the performance objectives of the curriculum set down for each grade from 2 to 8 in our system, by replacing items from the existing shelf test that did not measure our objectives at the appropriate grade level with items that did so and that matched the shelf test items they were to replace in terms of difficulty and level of discrimination. To this end, we made a list of objectives that were not measured by the levels of the existing shelf tests, and worked with the publisher to write potential replacement items that were then tried out in the spring, 1987 along with Form M of the shelf test.

At each grade about 200 "new" items were developed, items that clearly measured N.Y.C. curriculum objectives. (About half of these "new" items were items from other tests the publisher had available). All these items were arranged on tryout sheets

containing 20 items each, and were administered to samples of students along with Form M of MAT-6 in spring, 1987. A preliminary Rasch equating analysis was undertaken to provide information about the difficulty values of the new items. These data were then reviewed and a set of items were selected for the next year's customized tests to reflect both curriculum and psychometric considerations. Tentative estimates of difficulty values and of ability levels were generated for all possible total scores, and compared to those generated from the existing MAT-6 test.

We discovered that our plan to select tests that were perfect psychometric replicas of the shelf tests while also being perfect matches to our curriculum was not completely feasible. Several problems appeared to compromise our ideal design. First, while we developed and tried out items to cover the areas not represented by the shelf test, we did <u>not</u> do so with a view to proportional representation of items so that the test would reflect the emphases in the curriculum for given grades. That is, we did not attend enough at the initial stages to a "test blueprint". Rather, we developed an equal number of items to measure each objective that wasn't covered by the shelf test, not a number proportional to those that would be needed to fit a test blueprint that reflected the specific emphases of our curriculum. Second, for some content areas that were particularly important for our new and evolving curriculum, e.g., probability and statistics, item statistics were very poor, with low p and r

values. These items were nonetheless included at some levels since they reflected the heart of this new curriculum. Third, and perhaps most important, we were attempting to develop separate tests for each grade level, while the shelf test contains levels that span several grades. Our curriculum staff imposed an additional important constraint that the shelf test did not have. They did not want any item to appear on a test if the objective that item was measuring had not yet been introduced at that grade level. While on the one hand this makes a good deal of sense, it does not reflect the way achievement tests are built, particularly multi-grade test levels. Because of this, we found that we were severely limited in our ability to replace shelf test items with tryout items that differed only in their content emphasis, not psychometrically.

It is fair to say that, for item selection, curriculum interests were primary. Since these interests were the ones that spurred us on to customization, we felt that if we did not adequately satisfy these interests, we would be just as well off continuing to administer the shelf test. However, some compromises were made along the way by the curriculum staff members so that the resulting tests were close enough to the psychometric characteristics of the shelf test to maintain, we hope, both a norm-referenced interpretation and curriculum relevant usage.

Table 1 shows the average difficulty values for the tests developed for each grade, and the comparable MAT-6 shelf test

Table 1 A Comparison of Preliminary Scaled Difficulty Values for MAT-6 and Customized Tests for Each Grade

		Mean Scaled Difficulty Values	
		Customized Test	MAT-6
Total Math Scores			
Grade	2	511	510
	3	556	560
	4	569	560
	5	600	599
	6	614	599
	7	627	626
	8	636	626
Concepts			
Grade	2	530	521
	3	560	568
	4	574	568
	5	609	620
	6	626	620
	7	634	630
	8	647	630
Problem Solving			
Grade	2	502	503
	3	570	561
	4	580	561
	5	595	583
	6	608	583
	7	632	620
	8	635	620
Computation			
Grade	2	511	508
	3	556	549
	4	564	549
	5	600	589
	6	614	589
	7	627	620
	8	636	620

scaled difficulty values. The MAT-6 values are shown to be the same for Grades 3 and 4, for Grades 5 and 6, and for Grades 7 and 8, reflecting the use of Elementary, Intermediate and Advanced 1 levels for these pairs of grades, respectively. For the customized test, however, a unique test was designed for each grade level so that each grade has its own average difficulty value.

In examining Table 1, it is apparent that the tests developed for the lower grades are easier than those developed for the higher grades. This is certainly to be expected, and was comforting.

Another reasonable expectation of these data is that the items for the lower grade, within the pair of grades that use the same shelf test level, would be easier than the shelf test items and, conversely, items for the higher grade of the pair would be harder than the shelf test items. This would result in having a series of pairs, whose average difficulty would be the same as the shelf tests. For Grades 3 and 4, this was the case. For Grades 5 and 6, both tests are harder than the shelf test, but grade 6 is much harder, while grade 5 is only one point harder. The pattern for grades 7 and 8 mirrors that found for 5 and 6.

One would also expect that the pattern that is shown for Total Scores would be mirrored in that of the subtest scores. In Table 1, it can be seen that some differences in patterns of difficulty for the subtest scores, and their relation to shelf

test difficulty patterns, vary for the given grades. In Grade 3, for example, items for the Concepts subtest are easier than those for the Elementary Level of the shelf test. But, items for Computation and Problem Solving are harder than the shelf test. For Grade 4, all three subtests have items that are more difficult than the shelf test.

Thus, on the basis of information shown in Table 1, it appears that the "perfect match" of curriculum and psychometric qualities is made imperfect to varying degrees, depending on which subtest one is considering, and which grade level one examines. While imperfect, the scores do follow a logical progression from grade to grade, and do not deviate so significantly from the ideal pattern that one would suspect any serious compromise of the norms. The difficulty values shown in Table 1 are also to be considered with some caution since they include estimates for items that are included at one grade level, but that were actually tried out at another grade level. We will elaborate on the implications of this phenomenon later.

While our item selection is finished, our work is not yet complete. We face additional challenges as we move on to the final equating that must be done following this spring's test administration. As described previously, some items that were tried out at some of the lower grade levels were selected for inclusion on the upper grade level tests. This was done primarily to satisfy curriculum constraints and fill in

appropriate items for the test blueprints. While we expect the difficulty values to hold for items that were tried out and selected for inclusion at a given grade, we are concerned about the stability of difficulty estimates drawn from lower grade levels. This "jumping" of grade level, from the level it was tried out on to the level it was selected for, did not occur very frequently, but should be checked in a final equating.

At this point, one might ask whether it was all worth it. We found the project to be extremely costly and labor intensive, especially considering that decisions had to be reached with mutual agreement of the staff from two central N.Y.C. Board of Education Offices, (the Office of Educational Assessment and the Mathematics Unit in the Division of Curriculum and Instruction) and also of the staff from the publisher.

We do not know how the results will turn out, or how the tests will be viewed in the field and whether they will be used more extensively for instructional purposes. In a system that is as large and as politically sensitive as ours, we firmly hope that our efforts will produce reliable and valid information. We are certainly concerned that the results will appear in a way that people will in fact use them in as criterion-referenced a manner as possible, and that the normative information will show orderly progressions form grade to grade and be valuable in making norm-referenced interpretations. We believe our efforts were undertaken with as much care and consideration of the

legitimacy of a variety of needs and viewpoints from different sectors of our school as was possible. Despite the problems we tangled with, the logistics we struggled with and the compromises we made, in the long run we believe that our efforts will prove to have been worthwhile.